



# Gene Duplication and Fragmentation in the Zebra Finch Major Histocompatibility Complex

## Citation

Balakrishnan, Christopher N., Robert Ekblom, Martin Völker, Helena Westerdahl, Ricardo Godinez, Holly Kotkiewicz, David W. Burt, et al. 2010. Gene duplication and fragmentation in the zebra finch major histocompatibility complex. BMC Biology 8:29.

## Published Version

doi:10.1186/1741-7007-8-29

## Permanent link

<http://nrs.harvard.edu/urn-3:HUL.InstRepos:4454684>

## Terms of Use

This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Open Access Policy Articles, as set forth at <http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#OAP>

## Share Your Story

The Harvard community has made this article openly available.  
Please share how this access benefits you. [Submit a story](#).

[Accessibility](#)

RESEARCH ARTICLE

Open Access

# Gene duplication and fragmentation in the zebra finch major histocompatibility complex

Christopher N Balakrishnan<sup>1,8\*</sup>, Robert Ekblom<sup>2,3</sup>, Martin Völker<sup>4</sup>, Helena Westerdahl<sup>5</sup>, Ricardo Godinez<sup>1</sup>, Holly Kotkiewicz<sup>6</sup>, David W Burt<sup>7</sup>, Tina Graves<sup>6</sup>, Darren K Griffin<sup>4</sup>, Wesley C Warren<sup>6</sup>, Scott V Edwards<sup>1</sup>

## Abstract

**Background:** Due to its high polymorphism and importance for disease resistance, the major histocompatibility complex (MHC) has been an important focus of many vertebrate genome projects. Avian MHC organization is of particular interest because the chicken *Gallus gallus*, the avian species with the best characterized MHC, possesses a highly streamlined *minimal essential* MHC, which is linked to resistance against specific pathogens. It remains unclear the extent to which this organization describes the situation in other birds and whether it represents a derived or ancestral condition. The sequencing of the zebra finch *Taeniopygia guttata* genome, in combination with targeted bacterial artificial chromosome (BAC) sequencing, has allowed us to characterize an MHC from a highly divergent and diverse avian lineage, the passerines.

**Results:** The zebra finch MHC exhibits a complex structure and history involving gene duplication and fragmentation. The zebra finch MHC includes multiple Class I and Class II genes, some of which appear to be pseudogenes, and spans a much more extensive genomic region than the chicken MHC, as evidenced by the presence of MHC genes on each of seven BACs spanning 739 kb. Cytogenetic (FISH) evidence and the genome assembly itself place core MHC genes on as many as four chromosomes with TAP and Class I genes mapping to different chromosomes. MHC Class II regions are further characterized by high endogenous retroviral content. Lastly, we find strong evidence of selection acting on sites within passerine MHC Class I and Class II genes.

**Conclusion:** The zebra finch MHC differs markedly from that of the chicken, the only other bird species with a complete genome sequence. The apparent lack of synteny between TAP and the expressed MHC Class I locus is in fact reminiscent of a pattern seen in some mammalian lineages and may represent convergent evolution. Our analyses of the zebra finch MHC suggest a complex history involving chromosomal fission, gene duplication and translocation in the history of the MHC in birds, and highlight striking differences in MHC structure and organization among avian lineages.

## Background

The major histocompatibility complex (*MHC*) is a gene-dense genomic region within which many genes play a role in vertebrate immune response. *MHC Class I* genes encode surface receptors in most nucleated cell types and facilitate immune responses to intracellular pathogens. *MHC Class II* genes also encode receptors but are restricted to antigen presenting cells of the immune system where they play a role in combating extracellular pathogens. After the binding of antigens, Class I and Class II proteins present them to CD8 and CD4 T cells,

respectively. This presentation in turn triggers the adaptive immune response against the antigen. Polymorphism at *MHC* loci facilitates binding of a diversity of pathogens and this evolutionary selection pressure is thought to contribute to the high genetic variation in *MHC* loci [1]. *MHC* genes are perhaps the most thoroughly studied example of adaptive molecular evolution, representing a classic example of balancing selection [2-4]. *MHC* genes have also played an important role in studies of molecular ecology with *MHC* genotype influencing patterns of mate choice [reviewed in [5]], local adaptation [6], disease resistance [7], and the expression of sexually selected ornaments [8,9].

\* Correspondence: [cbala@igb.uiuc.edu](mailto:cbala@igb.uiuc.edu)

<sup>1</sup>Department of Organismic & Evolutionary Biology, Museum of Comparative Zoology, Harvard University, Cambridge, MA 02138, USA

An *MHC* has been identified in all jawed vertebrates studied to date. A core set of genes, including *TAP*, *TAPBP*, *TNXB* and *CENP-A*, are syntenic to the *MHC* across vertebrates that have been characterized [10-12]. Therefore, the presence of these genes, along with *Class I* and *Class II* genes can be used to define the *MHC* region. In addition to *Class I* and *Class II* genes, many of the other genes in the *MHC* region also play a role in immune response. Like *Class I* and *Class II* genes, *CD1* genes (which are *MHC*-linked in the chicken but not in mammals) play a role in the adaptive immune response. *CD1* molecules present lipid, glycolipid and lipopeptides to T and NKT cells [13]. *CD1* genes are in fact evolutionarily related to *Class I* and *Class II* genes [14-16]. Some *MHC*-linked genes encode proteins that interact with *MHC* molecules. *TAP1* and *TAP2* genes, for example, are involved in the loading of peptides onto *Class I* molecules for transport to the cell surface [17,18]. The interaction between *Class I* and *TAP* is itself mediated by *TAPBP* (tapasin).

While *MHCs* share many structural features there is also tremendous variation in their organization among species. Among birds, the chicken *Gallus gallus* has been most intensively studied, and its *MHC* (also known as *MHC-B* or B-complex) has a rather remarkable structure: a minimal essential *MHC* [19]. In contrast to humans, in which the *MHC* spans four megabases (MB) and consists of over 200 genes, the chicken *MHC* consists of only about 40 genes spanning only a few hundred kb on chromosome 16 [19,20]. In addition to the *MHC-B*, chicken *MHC Class I* and *Class II* genes are also present in a separate and unlinked cluster called the *MHC-Y* (or *rfp-Y*) region [21-25]. Even when considering both the *MHC-B* and *MHC-Y* together, the chicken *MHC* has fewer total genes, gene duplicates, pseudogenes, repetitive sequences, and shorter genes and introns than its mammalian counterparts.

Tight linkage of genes in the *MHC* may facilitate their coordinated coexpression [19,23,26] and coevolution in the chicken B-complex [27]. Furthermore, the suppression of recombination among *MHC* genes is thought to contribute to the evolution of gene complexes coadapted to particular pathogens and environments [26,28]. Interacting *TAP* and *Class I* genes are more closely linked in the chicken than in mammals and these genes in particular are thought to coevolve in birds [24]. Some of the strongest genotype/disease resistance correlations have been identified in the chicken [for example, [29,30]] and the simple architecture of the chicken *MHC*, with few highly expressed *MHC* genes, likely contributes to this pattern [19,23,24,26]. Due to the limited taxonomic and genomic sampling of *MHC* regions in birds, however, it remains unclear whether the streamlining of the *MHC* reflects the broader trend of reduced genome size in

birds [31,32] and whether a small *MHC* represents the ancestral condition for birds. Alternatively, a small *MHC* may be a highly derived condition unique to the biology of Galliforms.

The zebra finch genome, representing the taxonomically diverse Passerine clade (approximately 5,400 species), offers the opportunity to characterize *MHC* structure in an avian lineage highly divergent from the chicken. Molecular estimates of divergence between Passerines and Galliforms indicate that they diverged between about 90 and 120 million years ago [for example, [33]]. Among birds, only two Galliform *MHCs*, the domestic chicken and Japanese quail *Coturnix japonica*, have been well characterized [34,35]. The *MHC-B* complex of another Galliform species, the turkey *Meleagris gallopavo* has also recently been sequenced and appears similar to the chicken in structure spanning about 200 kb [36,37]. *MHC* polymorphism surveys in passerines suggest that their *MHC* may differ from the structure seen in Galliforms [7,38-40]. *Class IIB* genes in particular appear to have been extensively duplicated in passerine birds, although little is known regarding the expression of these genes. Initial attempts to characterize the passerine *MHC* regions using genomic sequence data have uncovered pseudogenes and have revealed a much lower gene density than the chicken [41-43]. The number of expressed *Class I* genes in some songbirds also appears greater than in the chicken [44]. None of the core *MHC*-associated genes described above have been characterized in passerines making it unclear whether classical *MHC* regions have been sequenced. In this study we used the draft assembly of the zebra finch genome [45] in combination with targeted BAC sequencing, fluorescence in situ hybridization (FISH) mapping, and restriction fragment length polymorphism (RFLP) analysis to describe the fundamental features of the zebra finch *MHC*.

## Results

### Genome assembly analysis

In our scan of the zebra finch genome assembly we found one or more homologous loci for 18 of 28 investigated chicken *MHC* related genes (Table 1). These represent whole coding sequence (cds) or fragments of genes (one or more exons). Since several of the genes we queried had multiple loci in the zebra finch assembly, our set of sequences comprises a total of 22 manually curated *MHC* genes and eight putative pseudogenes (sequences containing frame shift mutations or premature stop codons).

We found only one functional *MHC Class I* gene, which was situated on chromosome 22\_random (linked to chromosome 22, but exact location and orientation unknown). In addition, there are also 10 contigs

**Table 1 MHC genes identified in the survey of the zebra finch genome assembly.**

Gene	Chicken		Zebra Finch		Coordinates and orientation (+/-)	Ensembl ID
	# loci	Chr#	Locus ID	Chr#		
CD1	2	16	1	12	36510 – 39728 (+)	-
			2	12	31218 – 32904 (+)	ENSTGUG00000003538
TNXB	1	16		-	-	-
CYP21	1	16		22_random	5,200 – 8,483	ENSTGUG00000003380
CENP-A	1	16		Un	Contig 19574 (-)	ENSTGUG000000016809
C4	1	16		-	-	-
TAP1	1	16		14_random	8635 – 8114 (-)	ENSTGUG000000015337
TAP2	1	16		-	-	-
Class I	2	16	1	22_random	954 – 3750 (+)	ENSTGUG000000017273
(+MHC-Y)						
			ψ L	Un	Contig 5002 (-)	-
			ψ L	Un	Contig 19247 (+)	ENSTGUG000000016646
			ψ L	Un	Contig 43472 (-)	-
			ψ L	Un	Contig 237 (+)	-
			ψ L	Un	Contig 5814 (+)	-
			ψ L	Un	Contig 29268 (-)	ENSTGUG000000014179
			ψ L	Un	Contig 24227 (-)	ENSTGUG000000015460
			ψ L	Un	Contig 19531 (-)	-
			ψ L	Un	Contig 1325 (-)	ENSTGUG000000016290
			ψ C	Un	Contig 237 (+)	ENSTGUG000000015195
			ψ O	16_random	79190 – 79374 (+)	-
DMA	1	16		-	-	-
DMB	2	16		-	-	-
BRD2	1	16		Un	Contig 10922	-
Class IIB	2	16	1	Un	Contig 1486 (-)	ENSTGUG000000013745
(+MHC-Y)						
			1	Un	Contig 3597 (-)	ENSTGUG000000016075
			1	Un	Contig 12575 (-)	ENSTGUG000000015634
			1	Un	Contig 648 (+)	ENSTGUG000000014620
			1	Un	Contig 926 (+)	ENSTGUG000000017149
			1	Un	Contig 3052 (-)	ENSTGUG000000014503
			1	Un	Contig 11727 (+)	ENSTGUG000000015020
			2	Un	Contig 395 (+)	ENSTGUG000000016844
			2	Un	Contig 4424 (-)	ENSTGUG000000014905
			3	Un	Contig 2943 (+)	ENSTGUG000000014233
			4	Un	Contig 11297 (-)	ENSTGUG000000014649
			ψ L	Un	Contig 3510 (-)	ENSTGUG000000015846
			ψ M	22_random	279244-283106(+)	ENSTGUG000000017281
			ψ M	Un	Contig 3181 (-)	-
			ψ Q	7_random	92325-103474 (+)	ENSTGUG000000016701
			ψ R	Un	-	ENSTGUG000000017139
			ψ T	22_random	-	ENSTGUG000000017280
TAPBP	1	16		-	-	-
Blec1	1	16		Z	64162080 – 64162784 (+)	ENSTGUG000000005208
NKr	1	16		Z	64155557 – 64154733 (-)	-
BG	3	16		-	-	-
TRIM41	1	16		-	-	-
TRIM27	1	16		Z	64166328 – 64164434 (-)	ENSTGUG000000005203
TRIM39	1	16		Un	Contig 15508 (+)	ENSTGUG000000014157
TRIM27.2	1	16		-	-	-
TRIM7	1	16		16_random	756 – 1262 (+)	ENSTGUG000000015652

**Table 1: MHC genes identified in the survey of the zebra finch genome assembly. (Continued)**

LAO	1	16	16_random	127,342-132,827 (+)	ENSTGUG00000016298
TRIM7.2	1	16	16_random	147414 – 164815 (+)	ENSTGUG00000015672
KIFC1	1	16	22_random	7,798-8,481 (+)	-
Class IIA	1	16 Un	Un	Contig 28013 (+)	-
CIITA	1	14	14	6294135 – 6303566 (+)	ENSTGUG00000004838
li	1	13	13	6774603 – 6780145 (+)	ENSTGUG00000000882
B2M	1	10	10	-	ENSTGUG00000004607

Genes are listed in order according to their organisation on chicken chromosome 16. Duplicated genes are given a numbered Locus ID if they appear functional. Putative pseudogenes are marked with a  $\psi$  and given a lettered locus ID. Loci are classified based on sequence similarity such that identical sequences found in multiple places in the genome assembly are given the same Locus ID. Where Ensembl IDs have been assigned these are also given. Three genes outside of the MHC region but with a related function (CIITA, li, and B2M) are also included.

unincorporated into the genome assembly, and one contig on chromosome 16\_random, that contain fragments of *Class I* genes (Table 1). Sequence differences suggest that these genes correspond to at least three different pseudogenes. In the case of *MHC Class IIB*, we found 14 contigs on chromosome Un (unmapped genomic region), and three hits on chromosomes 22\_random and 7\_random, containing parts of the cds. Four distinct sequences with an open reading frame spanning exons two and three appear to be functional.

The genome assembly suggests that some *MHC*-associated genes may not be as clustered in the zebra finch *MHC* as they are in the chicken. *Blec1*, *NKR* and *TRIM27*, for example, map to the Z chromosome in the zebra finch genome assembly, while two *CD1* loci map to chromosome 12. Many of the other genes for which we searched, however, mapped to chromosome UN or were not found in the assembly (Table 1).

#### BAC screening, sequencing and gene prediction

We further characterized the zebra finch *MHC* by isolating and sequencing *MHC*-containing BAC clones. We first identified 96 clones that hybridized strongly with a probe targeted to exon 3 of an *MHC Class IIB* gene. Four of these BACs were selected for sequencing (hereafter *Class II* clones). Because of the large number of *Class IIB* positive clones, we conducted further screening using overgo probes targeted to five conserved genes linked to the *MHC* across a diversity of taxa (Table 2; Additional File 1). For this second screening, we

**Table 2 Results of overgo hybridization of zebra finch BAC library.**

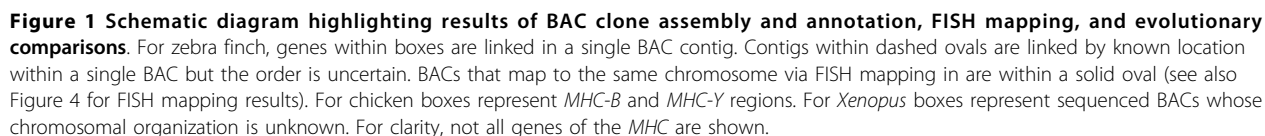
	KIFC	Class I	TNXX	TAP2	CENP-A
Positive Clones	56	28	11	14	44
Co-hybridizing clones	16	1	4		

Cohybridization patterns link all genes but *CENP-A*. The clone that was positive for *TNXX* and *Class I* appears to contain a *Class I* pseudogene.

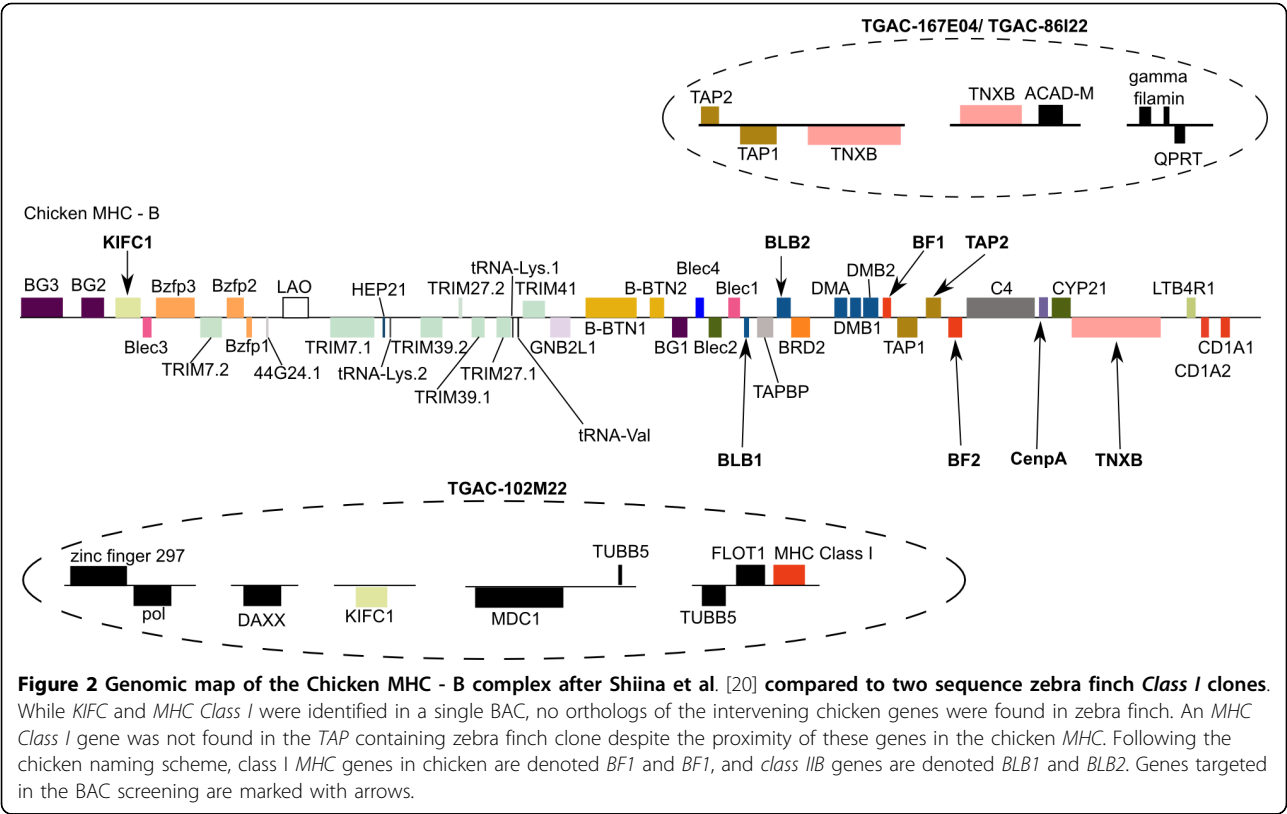
screened a different BAC library derived from the same zebra finch individual as the whole genome sequence (see methods). Positive clones were found for each of the five genes: *MHC Class I* (n = 21), *KIFC* (n = 56), *CENP-A* (n = 44), *TAP2* (n = 14), and *TNXX* (n = 11). Probes for three pairs of genes were found to cohybridize to individual BAC clones: *MHC Class I* and *KIFC*, *MHC Class I* and *TNXX*, and *TNXX* and *TAP2* (Table 2). One BAC clone containing each of these three gene pairs was chosen for 6× sequencing (hereafter *Class I clones*). No clones were positive for both *TAP2* and *MHC Class I*, suggesting that these two genes are not closely linked in the zebra finch as they are in the chicken. *CENP-A* probes also did not cohybridize with any of the other *MHC* genes, again indicating a lack of close linkage observed in other species (Figures 1 and 2). We did not sequence any of the *CENP-A* positive clones.

Each clone was assembled into multiple ordered contigs (Table 3). The fragmented nature of the BAC assemblies is expected given the coverage, but was exacerbated by high repeat content (see below; Additional File 2). Sequence analysis of two *Class I* clones, *TGAC-86I22* and *TGAC-167E04*, revealed extensive sequence overlap and thus were assembled together (Table 3). To improve the assembly for the *Class II* clones we generated additional sequencing reads. Because the *Class I* clones were derived from the same zebra finch as the whole genome sequence, we were also able to incorporate sequence reads from the whole genome sequencing effort to improve the assembly of these BACs.

Most of the BAC assemblies contained at least a fragment of the genes expected based on the probes used to identify them (Additional File 3). *Class I* clones contained a number of genes of interest. The assembly of clones *TGAC-86I22* and *TGAC-167E04* contained sequences with similarity to *TAP1*, *TAP2*, and *TNXX* genes, and clone *TGAC-102M22* contained *KIFC*, *DAXX*, *TUBB*, *Class I*, and *FLOT* (Figure 1). Although numerous genes separate *KIFC* and *Class I* genes in chicken (Figure 2) we did not find evidence for these



sequencing only revealed a small region with similarity to the *Class I* 3' UTR in the great reed warbler (e-value: 8e-19, identities: 166/230; 72%). Polymerase chain reaction (PCR) screening of this clone also identified a stretch of a short exon 3 sequence (200 bp) that is identical to the expressed locus. The UTR region, however, is distinctive in sequence from the expressed zebra finch



*Class I* 3' UTR and BLAST searches of brain expressed sequence tags (ESTs) and 454 sequencing data from multiple tissues suggest that this locus is not expressed [46]. Together this suggests that clone *TGAC-86I22* contains an *MHC Class I* pseudogene.

*Class II* clones contained numerous predicted genes with sequence similarity to zinc finger genes, as well as *gag* and *pol* proteins (endogenous retroviral genes). Aside from these and the expected *Class IIB* sequences, however, only one other gene of interest was found. A gene whose best blast hit matched the first four exons of the turkey *TAPBP* gene (blastx e-value 6e-22) was found in clone *TGAA-157B03*. Sequence conservation mapping using Zpicture [47] of this clone and a previously sequence red winged blackbird *Agelaius*

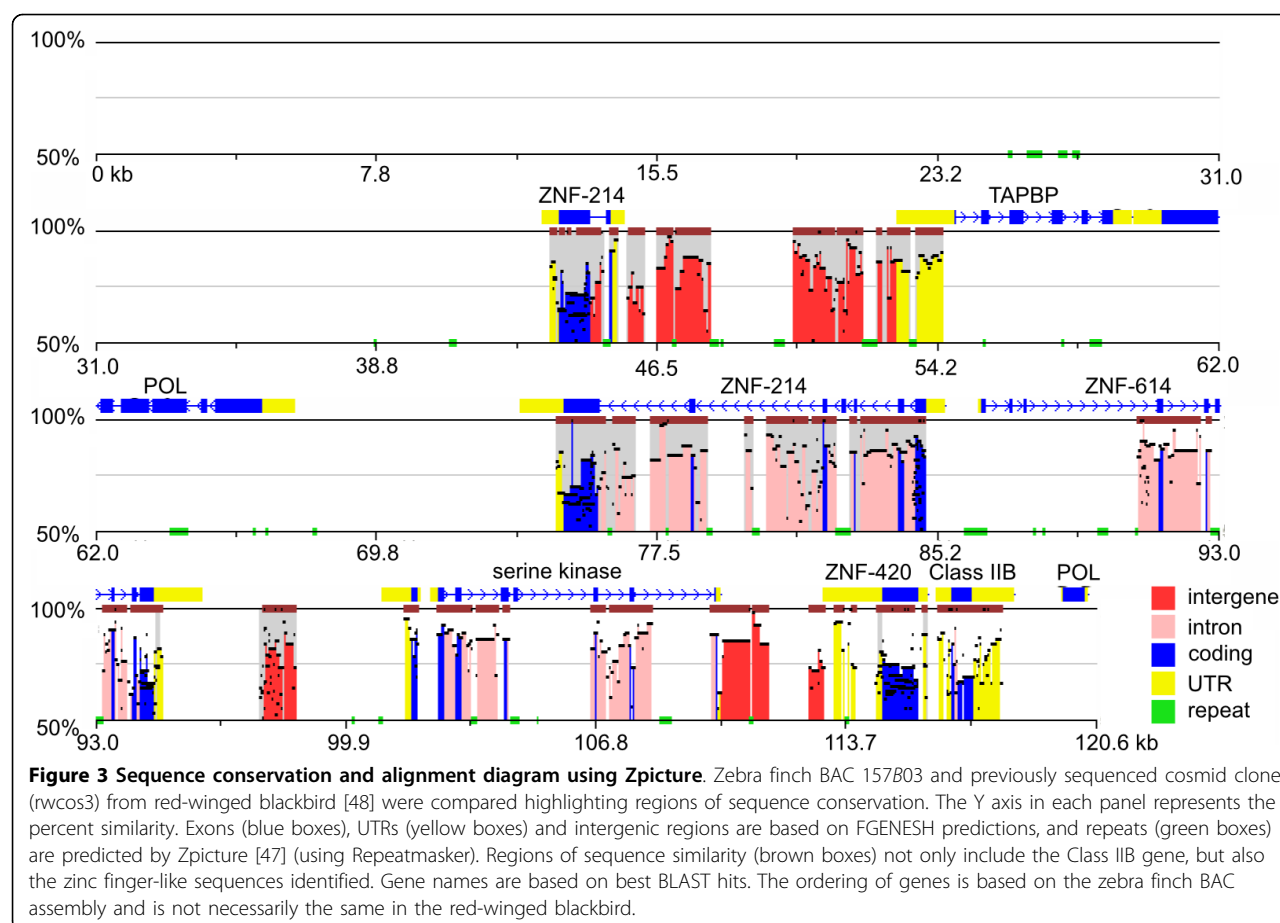
*phoenicius* *Class II* region [48], highlight sequence similarities in the coding and UTRs of predicted genes, as well as in some putative intergenic regions (Figure 3).

**FISH mapping of BAC clones to zebra finch chromosomes**

Single-color FISH mapping experiments revealed that sequenced *Class II* BACs (*TGAA-157B03*, *TGAA-351E14*, *TGAA-323J16* and *TGAA-47O03*) hybridized to several pairs of microchromosomes each (Figure 4), likely due to the high repeat content in these clones (see below). In contrast, sequenced *Class I* BACs (*TGAC-86I22*, *TGAC-102M22* and *TGAC-167E04*) each hybridized to one pair of small microchromosomes. Some BACs also cross-hybridized to repeats in the centromeric and telomeric regions of macrochromosomes (for

Table 3 Description of BAC assemblies.

BAC Clone	GenBank #	# contigs	largest contig	average length	total length
TGAA-157B03	AC192433	2	79,546	60,262	120,523
TGAA-323J16	AC191651	4	75,043	32,336	129,342
TGAA-351E14	AC191861	3	91,199	45,792	137,376
TGAA-047O03	AC192431	3	58,397	38,629	115,189
TGAC-102M22	AC232985	12	20,620	9,274	111,298
TGAC-167E04/TGAC-86I22	AC232854	17	25,067	7,355	125,027
total assembled length:					738,755



example, Figure 4). Linkage analysis by dual color FISH demonstrated that BACs *TGAC-102M22* (containing presumptive *Class I*, *FLOT*, *TUBB*, *KIFC*, *DAXX*), and three *Class II* BACs (*TGAC-323J16*, *TGAC-351E14* and *TGAC-470O3*) shared hybridization to one pair of small microchromosomes. Sequenced BACs *TGAC-86I22* and *TGAC-167E04* (containing presumptive *TNXB*, *TAP1*, *TAP2*) and *157B03* (*Class II*, *TAPBP*), however, hybridized to a different pair of small microchromosomes (Figure 4). *MHC* genes are thus found in two linkage groups on separate chromosomes in zebra finch.

To further test whether *TAP* and *MHC Class I* genes are syntenic, we conducted five additional two-color FISH experiments with BAC clones that were positive for *TAP2* and *MHC Class I*. While some *MHC Class I* probes hybridized to multiple microchromosomes, in only one case did we find colocalisation of *Class I* and *TAP2* probes (Table 4, Additional File 4). In this case, *MHC Class I* probes hybridized to multiple microchromosomes, and the colocalisation occurred on the W sex chromosome. It is therefore likely that this colocalisation is due to nonspecific binding, and the repetitive nature of the avian W chromosome. In total we have

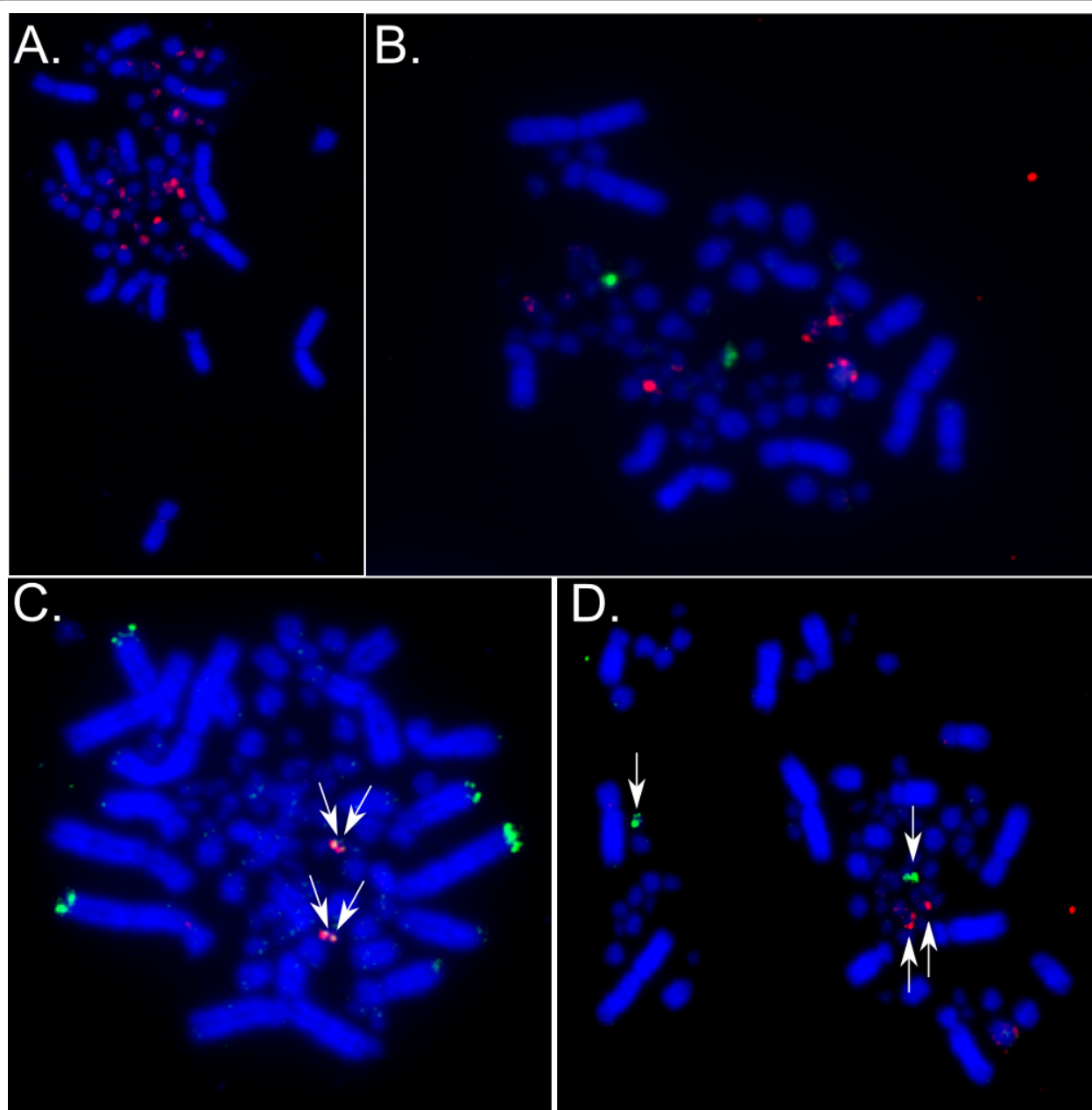
four cases in which *TAP2* probes hybridize unambiguously to a single microchromosome and in all of these, *Class I* maps to a different chromosome.

In order to identify the zebra finch chromosomes corresponding to the two zebra finch linkage groups, we performed dual-color FISH experiments in which one BAC from one of the two linkage groups (*TGAC-102M22* or *TGAC-86I22*) was co-hybridized with a non-MHC BAC with known chromosomal location (Additional File 5). These experiments covered all microchromosomes for which BACs are currently available (chromosomes 9 to 15 and 17 to 28). Neither of the two *MHC* linkage groups mapped to these chromosomes, suggesting that both microchromosomes to which the *MHC* BACs mapped may indeed correspond to parts of zebra finch chromosome 16, the only chromosome for which we do not have known BACs.

#### Polymorphism survey via RFLP/Southern Blot

To test our findings on the numbers of *Class I* and *Class IIB* genes, and as a preliminary survey of gene number and intraspecific polymorphism, we conducted a RFLP analysis using probes targeted to these loci.





**Figure 4 FISH mapping of BAC clones.** **A)** Single color FISH mapping of TGAC-157B03 reveals extensive cross-hybridization across chromosomes. Similar results were observed for other Class II clones presumably as a result of their high repeat content. **B)** Lack of cohybridization between Clones TGAC-102M22 and a known chromosome 22 BAC indicates that TGAC-102M22 is not on chromosome 22 as indicated by the genome assembly. **C)** Dual color FISH of TGAC-86I22 (red) and TGAC-167E04 (green) indicating cohybridization of these clones, a result also supported by sequence analysis. These clones were assembled together, and contain g-filamin, TNXB, TAP1 and TAP2 genes. **D)** Clones TGAC-102M22 (red) (contains MHC Class I, FLOT, TUBB, KIFC and DAXX) and TGAC-86I22 map to different chromosomes. Key components of the classical MHC therefore map to different chromosomes in the zebra finch genome.

There are clearly a larger number of RFLP bands for *MHC Class IIB* (range = 12 to 27) than for *Class I* (range = 2 to 4) and this also suggests that there are more *Class IIB* genes than *Class I* genes in zebra finches (Figure 5). This difference in gene number is not likely due to differences in the sequence similarity of probe

and target for *Class I* and *IIB* probes as we would expect that the longer *Class I* probe (280 bp) should hybridize to a larger number of fragments than the *Class IIB* probe (207 bp). We repeated this hybridization twice using different *Class I* and *II* probes (data not shown). For *MHC Class I* there are two to four RFLP

**Table 4 Two-color FISH mapping results of putative TAP 2 and MHC Class I-containing clones.**

MHC Class I	Mapping	TAP 2	Mapping	Colocalisation
TGAC-102M22*	1 micro	TGAC-167E04*	1 micro	No
TGAC-102M22*	1 micro	TGAC-86I22*	1 micro	No
TGAC-15A11	1 micro	TGAC-95I13	centromere macro	No
TGAC-181L18	1 micro	TGAC-53B12	Z	No
TGAC-12A09	micros, W	TGAC-14G17	1 micro and W	Yes
TGAC-250C06	micros	TGAC-139M05	1 micro	No
TGAC-252P06	1 micro	TGAC-249G24	1 micro	No

\* Clones sequenced in this study

Most clones mapped to a single microchromosome (1 micro) where as some showed non specific binding to multiple microchromosomes (micros). A few clones mapped to sex chromosomes (Z and W), and one mapped to the centromeric region of a macrochromosome (macro). In only one cases did TAP2 and Class I clones colocalise, and that was on the W chromosome.

fragments in the captive zebra finches from the US (ind 1 to 7) and two to three fragments in the zebra finches from Sweden (ind 8 to 10). For *MHC Class IIB* there are 12 to 20 RFLP fragments in the zebra finches from the US and as many as 27 fragments in the three zebra finches from Sweden.

#### Comparative analysis of MHC genes

In order to explore the evolution of the compact avian *MHC* structure observed in chicken, we estimated the average gene density in quail, chicken, zebra finch and human. Interestingly, the estimated gene density in the zebra finch is similar to that in humans, and distinctive from both quail and chicken (Figure 6). Repeat content also differs markedly between the chicken *MHC* region and the zebra finch BACs. Long interspersed nuclear elements (LINEs) occur at frequency of 0.07 per kb in chicken versus 0.02 per kb across the 739 kb of zebra finch BAC sequence. Long terminal repeat (LTR) content, specifically in the form of *ERV1* elements, was exceptionally high in zebra finch *Class II* sequences, occurring at frequency of 0.14 per kb, whereas the chicken *MHC* is depauperate in LTR at 0.01 per kb (Figure 7). Zebra finch *Class I* clones sequenced here more closely resemble the chicken content, but still had higher LTR content (0.02 LTR/kb).

To examine the evolutionary relationships among *MHC* genes, we placed exons 2 and 3 of four putatively functional zebra finch *MHC Class IIB* sequences in a phylogenetic context by comparing them with other passerine sequences from GenBank. Exon 2 of the *Class IIB* gene encodes the protein that forms the

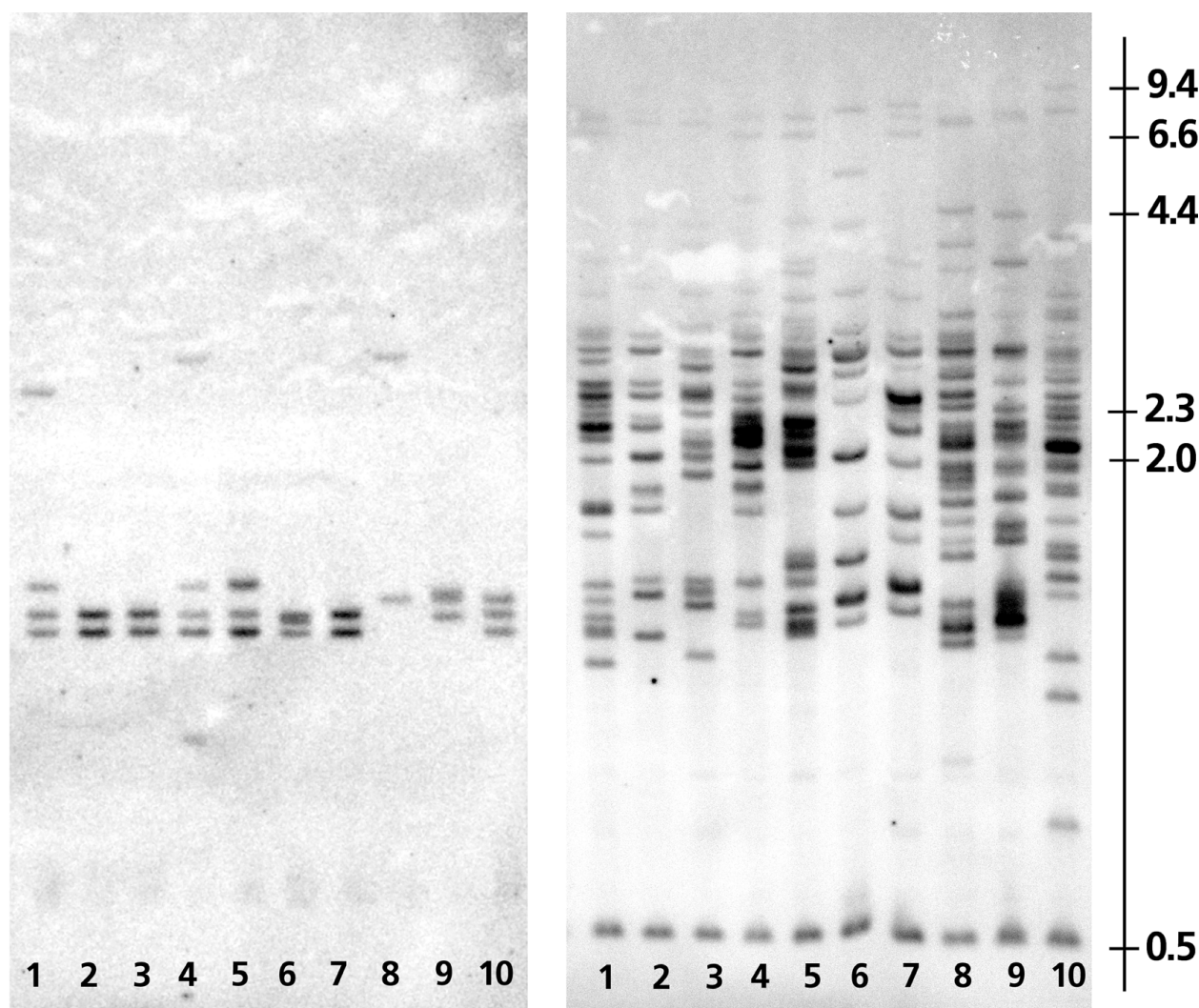
peptide binding region of the *Class II* protein, and exons 2 and 3 have been amplified and sequenced in a diversity of bird taxa. Bayesian phylogenetic analyses using both exons concatenated together grouped three zebra finch sequences in a highly supported clade (Figure 8). Using outgroups to the passerine sequences (data not shown), a divergent zebra finch locus was placed basal to all passerine sequences, between passerine and non-passerine sequences. Thus for analyses of passerine sequences we rooted the tree at this zebra finch sequence (Figure 8). We also analyzed exons 2 and 3 separately. These results reflect previously described differences among the exons [for example, [38]] so are not described further here. Phylogenetic analyses of exon 3 sequences from *MHC Class I* also placed zebra finch *Class I* sequences in a strongly supported clade (Additional File 6).

Passerine sequences were analyzed using Phylogenetic Analysis using Maximum Likelihood (PAML) [49,50], to test for the influence of positive selection on *MHC Class I* exon 3, and *Class IIB* exon 2. Two pairs of models were compared in CODEML, M1 (nearly neutral) versus M2 (positive selection) and M7 ( $\beta$ ) versus M8 ( $\beta$  and  $\omega > 1$ ) [49,50]. For *Class I* exons both tests indicate that models incorporating positive selection provide a better fit to the data than do neutral models (M1 vs. M2,  $2\Delta\ln L = 270.5$ ,  $P = 0.00$ ; M7 vs. M8,  $2\Delta\ln L = 271.7$ ,  $P = 0.00$ ). Bayes Empirical Bayes analyses [49] identified 10 sites with  $\omega > 1$  (posterior probability  $> 0.95$ ; see Additional File 6). Analyses of *MHC Class IIB* exon 2 also suggest a strong influence of positive selection (M1 vs M2,  $2\Delta\ln L = 540.4$ ,  $P = 0.00$ ; M7 vs M8,  $2\Delta\ln L = 511.0$ ,  $P = 0.00$ ) with 21 sites with  $\omega > 1$  (Figure 8). In both *Class I* and *Class IIB*, these sites correspond well with the peptide binding region (PBR) from the human *MHC* [51,52] and with selected sites identified in birds of prey [53,54].

#### Discussion

We have provided here a detailed characterization of the zebra finch *MHC*. There is clear cytogenetic evidence that *MHC* genes map to at least two different chromosome pairs in the zebra finch. If the chicken *MHC* represents the ancestral state, the situation in the zebra finch may have arisen through fission of chromosome 16 or a translocation of part of it to another pair of microchromosomes. The hypothesis of chromosomal fission is consistent with the finding that the *MHC* BACs did not map to zebra finch chromosomes 9 to 15 or 17 to 28, and that the microchromosomes recognized by these probes were small.

The finding of *MHC* genes on two chromosomes in the zebra finch is particularly intriguing because *TAP* genes map to one of them, whereas an expressed *Class I*

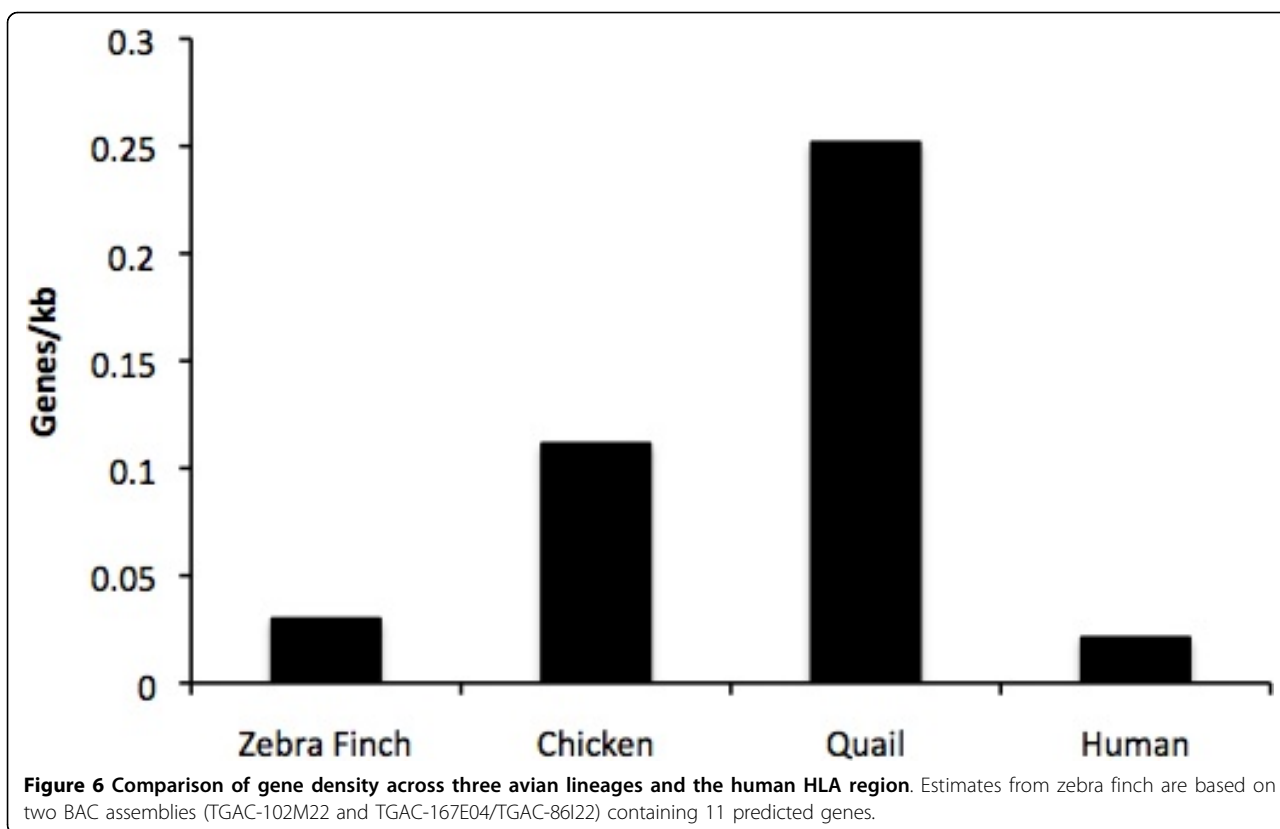


**Figure 5 RFLP/Southern Blot of 10 captive zebra finches.** Individuals 1 to 7 are from a captive American population and individuals 8 to 10 are from a Swedish population. The left panel shows the banding patterns using a Class I probe and the right panel shows the results using a Class II probe. Results from Class I analysis suggest a minimum of two loci whereas Class II probes indicate a very large number of loci (mean number of bands =  $19 \pm 4.6$ , range: 12 to 27).

gene (and a number of other *MHC*-associated genes) maps to a distinct chromosome (Figure 1). This finding is unexpected because *TAP* and *Class I* genes functionally interact and are syntenic in most MHCs studied to date including both chicken and humans [reviewed in [10], but see [55,56]]. In chicken this tight linkage is thought to result in coevolution between *TAP* and *Class I* genes and strong correlations between *MHC* haplotype and disease resistance [reviewed in [57]]. *TAP* genes in mammals, while generally syntenic, are not as closely linked to *Class I* as they are in Galliform birds. The separation of *TAP* and *Class I* in mammals has been hypothesized to have resulted in their evolutionary independence and in turn led to high levels of duplication and divergence in *Class I* genes [56]. This dissociation is

perhaps most clearly illustrated by the tammar wallaby *Macropus eugenii* in which *Class I* sequences have been found dispersed across seven chromosomes [56]. The separation of *TAP* and *Class I* genes in the zebra finch may therefore represent convergent dissociation of these genes.

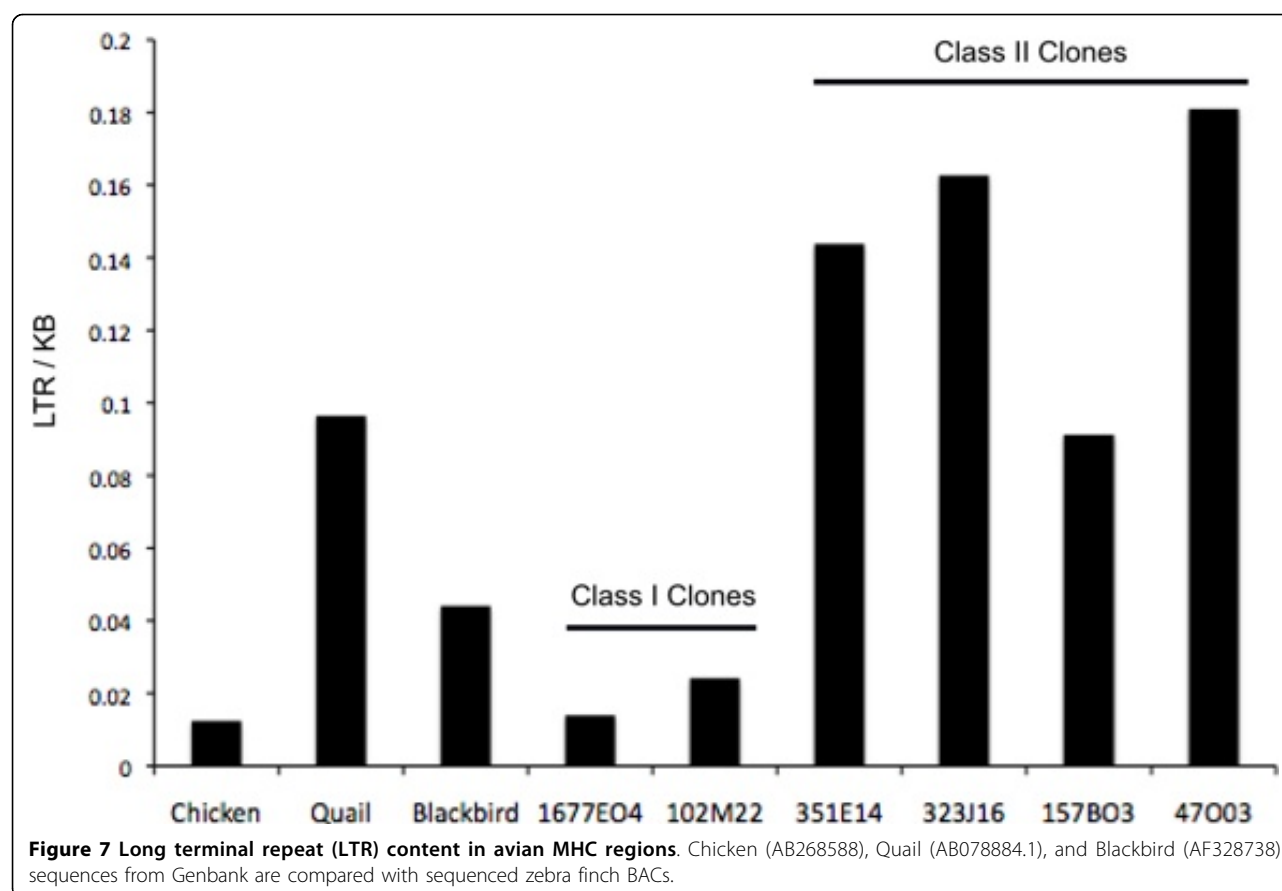
An alternative explanation for the separation of *Class I* and *TAP* genes in the zebra finch is that the regions sequenced here could represent duplication blocks. The sequenced *Class I* locus could even be related to the *MHC-Y* region of chicken. Phylogenetic analyses of zebra finch *Class I* and chicken *Class I* (*MHC-B* and *MHC-Y*), however, suggest that our sequenced *Class I* gene is not the ortholog of a chicken *MHC-Y* gene as chicken (including *MHC-Y*) and zebra finch sequences



are reciprocally monophyletic (Additional file 6). It is also possible that a second Class I gene resides on the same chromosome as *TAP1* and *TAP2* and therefore, that Class I and *TAP* are actually syntenic. In fact, a sequenced BAC was positive for both *MHC Class I* and *TNXB*; Another four clones were positive for *TNXB* and *TAP2* suggesting a possible linkage between these *MHC Class I* and *TAP2*. Based on a divergent sequence and a lack of expression, we suggest that this *Class I* sequence is a pseudogene. Even if it were not a pseudogene, *TAP* and this *Class I* gene would be much more distantly located in zebra finch than they are in chicken and would be free of the linkage seen in the chicken. The whole genome assembly, digital expression profiling [46] and EST data suggest only one full-length, expressed, *Class I* gene. It is also possible that there is a second set of *TAP* genes that we have not sequenced. Given the extremely low coverage of *TAP* genes in the genome trace archives (for example, only one read covering *TAP2*), it is unlikely that *TAP* genes have been duplicated. FISH mapping of five pairs of putative *TAP2* and *MHC Class I* clones further supports the lack of synteny among *TAP* and *Class I* genes (Table 4). Together these findings suggest that the *Class I* and *TAP* are not linked in the zebra finch. In addition to *Class I* loci identified in the BACs, we identified three distinct *Class I*

sequences that appear to be pseudogenes. One of the putative pseudogenes only contains exon 2, one only contain exons 4 to 6, while the third contains exons 1 to 3. Because the probes used in RFLP analyses target exon 3 only one of these pseudogenes would be reflected in the RFLP banding patterns. While the zebra finch appears to possess only one expressed *Class I* locus, the great reed warbler *Acrocephalus arundinaceus*, another passerine species, expresses multiple *Class I* loci [44]. An intriguing possibility is that the dissociation of *TAP* and *Class I* in ancestral passerines preceded the radiation of *Class I* genes in some passerine groups [44] as has been suggested for the wallaby [55,56].

*Class IIB* genes in zebra finch are highly duplicated as evidenced by the genome assembly, BAC sequencing and the RFLP analysis. We identified 10 distinct *Class IIB* sequences in the genome assembly (Table 1) some of which appear to be pseudogenes. These findings corroborate previous surveys of *Class IIB* variation in other passerine birds [40,58,59]. Another feature of zebra finch *Class IIB* regions is their high LTR content, mostly in the form of ERV elements (Figure 7). The finding of multiple zinc-finger genes and retroelements in proximity to *Class II* genes was also presaged by multikilobase *MHC* sequences from red-winged blackbirds, which showed a similar pattern [43,48]. Given the large

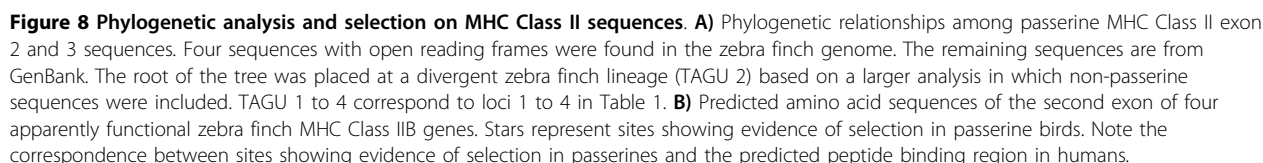


number of *Class IIB* duplicates and pseudogenes we speculate that duplication may have been related to the presence of retroviral sequences. Thus, the passerine *MHC Class IIB* may have been invaded by endogenous retroviruses much like the primate *Class I* [28]. Endogenous retroviruses have also been implicated in the duplication of wallaby *Class I* genes and their spread across multiple chromosomes [56].

Given the FISH mapping results and the whole genome assembly, *MHC* genes appear to be located on even more than two chromosomes. The genome assembly suggests that homologs of chicken *MHC* genes have been dispersed in the genome. There are at least three possible explanations for this: 1) There have been chromosome rearrangements for these genes between the chicken and zebra finch; 2) The contigs containing these genes have been misplaced in either the chicken or the zebra finch genome assembly; 3) The zebra finch gene identified is not the true ortholog of the chicken gene. Chicken *MHC* genes placed on different chromosomes in the zebra finch assembly compared to the chicken include *MHC Class I* (*Chr22\_random*), *CD1* and *CD2* (*Chr12*), and *NKR*, *Blec1* and *TRIM27* (*ChrZ*) (Table 1). The *MHC Class I*

gene placed on chromosome 22 and its surrounding region in the assembly is essentially identical to that in our sequenced BAC. This sequenced BAC did not cohybridize with two known chromosome 22 BACs (Figure 4B; Additional file 5), so the placement of this *Class I* region on chromosome 22 appears to be an assembly artifact. Rather, the FISH mapping results suggest that these genes are in fact on chromosome 16 as they are in chicken. The genome assembly data underlying the placement of *CD1* genes on chromosome 12 is also somewhat uncertain, with no BAC-end sequences linking contigs containing these genes to chromosome 12. Further work will be needed to test whether the genome assembly has properly placed these genes. Contigs containing, *Blec1*, *NKr* and *TRIM27*, however, are linked by BAC-end sequence pairs to the Z chromosome, making it likely that these are appropriately placed in the assembly.

A number of core *MHC*-associated genes including *DMA*, *BG*, *C4*, *TNXB*, *TAP2* and *TAPBP* are conspicuous by their absence in the zebra finch genome assembly (Table 1). There is no reason, however, to believe that these are truly absent in the zebra finch as they are present in a wide range of other vertebrates and are



BAC sequencing revealed two genes, *FLOT* and *DAXX*, that are MHC-linked in non-avian vertebrates [10,11], but have not been described in chicken. The relatively close linkage to MHC Class I and II genes of *FLOT*, *TUBB* and *DAXX* in the zebra finch is actually more similar to the organisation in some teleost MHCs [for example, [12]] than it is to either *Xenopus* or the human MHC, where *DAXX* is physically distant from



the *FLOT* and *TUBB* genes. Chicken chromosome 16, like the zebra finch, is not well assembled at this point so it is possible that these genes will be found as the chicken assembly continues to improve.

Phylogenetic analyses highlight the clustering of Class IIB loci by species rather than by orthology relationships, suggesting a history of concerted evolution, at least on portions of the genes [38,60,61]. We did, however, identify a unique Class IIB lineage that falls at the base of all other passerine Class II sequences. This appears to be a novel locus that has not previously been sequenced in birds and it is unknown whether it is expressed and/or polymorphic. Further analysis will be needed to clarify the role of this locus but its discovery underscores the utility of genomic approaches (rather than PCR amplification using degenerate primers) for characterizing MHC genes in birds. Tests of selection using zebra finch and other passerine MHC sequences support a strong role of selection in shaping patterns of polymorphism in the peptide binding region of Class I and Class II genes in passerines. The specific sites under positive selection are similar to those previously identified for other bird groups [53,54] and they closely match the peptide binding regions in humans [51,52]. High variability among individuals in RFLP banding patterns support the prediction that MHC Class IIB genes are influenced by balancing selection.

Among birds, there is tremendous variation among lineages in the number of MHC genes. In quail [34], red-winged blackbird [42,48] and the zebra finch, there are multiple Class II genes. Most non-passerine species, in contrast, appear to have only between one and three loci [60,62,63]. Given the derived phylogenetic position of passerines [64], these patterns imply that in terms of Class II genes, a minimal MHC may be ancestral for birds [60,62]. Because of the extensive variation among avian lineages in the number of Class I genes [for example, [34,44,65]], it remains unclear what the ancestral condition for Class I genes might be.

## Conclusions

We have made significant progress towards the understanding of the complex structure of the zebra finch MHC, the first such analysis from a representative of the diverse passerine radiation. Although the genome assembly and BAC sequencing are fragmentary, the zebra finch appears to possess an MHC differing markedly from previously described avian MHCs. The genomic architecture of the zebra finch MHC highlights the dynamic nature of MHC evolution. The evidence for gene duplication, pseudogenization and the distribution of MHC genes on multiple chromosomes in the zebra finch are particularly striking when measured against the compact MHC of the chicken present on a single

chicken microchromosome. Further genomic characterization of MHCs from a broader diversity of birds, as well as further refinement of the zebra finch MHC assembly, will continue to refine our picture of MHC evolution in birds.

## Methods

### Genome assembly scan

We searched the zebra finch genome assembly extensively for MHC genes using a variety of methods. Chicken MHC genes and proteins were downloaded from the National Center for Biotechnology Information (NCBI) website and blasted (blastn and tblastn) against the published version of the zebra finch genome and the available EST library and MHC containing BACs (see below). Since many of the genes of the MHC are diverging quickly we used rather relaxed blast settings (high minimum e-value and low w). For especially tricky multi-gene families and genes not found using the regular blast searches we constructed alignments using several vertebrate species and searched using conserved regions only. We also constructed a hidden Markov model of conserved features using the program HMMER 2.3.2 [66] and used the output consensus sequence in an additional blast search. The HMMER model was also used with the program Wise2 [67] in an additional attempt to identify corresponding exons in the zebra finch genome.

Regions in the zebra finch genome with significant hits on one or more chicken MHC exons were aligned to each of the chicken exons from the target gene using ClustalW [68] and checked manually in BioEdit [69]. Zebra finch sequences matching chicken MHC exons were extracted and complete or partial coding sequences of genes were blasted (blastx) back against the chicken RefSeq protein database. Hits with a best reciprocal blast with an e-value of less than 1e-05 against the target gene in chicken were considered to be orthologs.

Most of the genes were also identified using automated annotation of the zebra finch genome. In these cases we have included the accession numbers for the ENSEMBL entries (Table 1). These results, however, were not available to us at the time we conducted our analysis and have not affected our gene finding. Instead our manual annotation provides support for many of the genes identified using the computerised ENSEMBL annotation [70]. Also note that in some cases there are slight differences between the sequences presented here and the sequences with the provided ENSEMBL IDs.

### BAC screening, sequencing and gene prediction

We characterized the zebra finch MHC by isolating and sequencing MHC-containing BAC clones. To generate a probe for MHC ClassIIB we PCR amplified exon 3 using degenerate primers described by Edwards et al.

[71]. Probes were then radioactively labeled and hybridized with eight BAC filters following previously described protocols [72]. BAC filters were purchased from the Arizona Genomics Institute <http://www.genome.arizona.edu>. Clones from the AGI library are listed by names with the *TGAA* prefix. Positive clones were fingerprinted and four clones representing two pairs of putatively overlapping clones were sequenced to 6× coverage using an Applied Biosystems 3730 sequencer (Foster City, CA, USA). Because MHC Class IIB clones had apparently high repeat content, additional plasmid end reads were generated to improve the assembly.

We conducted additional screening using oligonucleotide probes targeted to five conserved genes linked to the MHC across a diversity of taxa (Table 2). This second round of screening was conducted using a different BAC library (TGAC), available through the Clemson University Genome Institute because this library was generated using DNA from the same individual zebra finch as the genome itself. Screening was done using previously established protocols [73]. Oligos were typically 24-mers (Supplementary Table 1) overlapping by 8 bp to generate a radiolabeled double-stranded 40-mer. These 40-mers were then pooled by gene and hybridized against the filters to identify BACs containing the specific gene sequence. Once the clones were identified, they were fingerprinted and end sequenced to confirm their location in the region. For these BACs we incorporated overlapping reads from the genome into the final BAC assemblies to increase contig length and improve ordering. Individual BAC assemblies were created with PHRAP [74] and assessed for contiguity. BAC assemblies were then manually examined for misassemblies and if they were found, the data was sorted as best as possible by using forward and reverse pair data. Consensus sequence blocks for each clone were then ordered and subjected to further analysis.

We used FGENESH [75] and GENSCAN [76] to predict genes contained within the BAC sequences. FGENESH uses a hidden Markov model (HMM) for gene prediction, and we used both human and chicken databases for gene prediction. Predicted amino acid sequences were blasted (blastp) against the non-redundant protein database in Genbank. Predicted genes with strong blast hits were given putative gene names, and were visually inspected to further confirm orthology with known genes. We assessed repeat content of clones using RepeatMasker [77] and chicken repeat libraries.

#### FISH mapping

Zebra finch chromosome preparations were made as previously described [78]. BACs were isolated using the Qiagen Plasmid Midi Kit (Crawley, UK). A total of 500 ng of isolated BAC DNA were labelled with

biotin-16-dUTP or digoxigenin-11-dUTP (Roche Applied Science, Burgess Hill, UK) by nick translation and resuspended in 10 µl of hybridization buffer (50% formamide, 20% dextrane sulphate in 2×SSC). Slides with metaphase chromosomes were dehydrated in an ethanol series (70%, 80%, 100%, three minutes each), aged for one hour at 75°C and treated with RNase A (100 µg/ml in 2×SSC) for one hour at 37°C. Chromosomes were denatured for 90 seconds in 70% formamide in 2×SSC at 75°C. Labelled BACs were mixed with hybridization buffer and chicken genomic DNA or herring sperm DNA (Sigma-Aldrich Company Ltd., Dorset, UK) in a 1:2:1 ratio, applied to slides and sealed under cover slips. Hybridization was carried out in a humidified chamber at 37°C (overnight for same-species hybridizations and for three days for cross-species hybridizations (see below)). Post-hybridization washes for same-species hybridizations consisted of 50% formamide in 2 × SSC for 2 × 10 minutes at 37°C; one minute in 2 × SSC with 0.1% Igelal at RT; 15 minutes in 4 × SSC with 0.05% Igelal at RT; 25 minutes in 4 × SSC with 0.05% Igelal and 2% BSA at RT. For cross-species hybridizations, the first washing step was modified (10% formamide in 2 × SSC for 2 × 10 minutes at 30°C). Probes were detected with 1:200 streptavidin-Cy3 (Amersham, Little Chalfont, UK), in 4 × SSC, 0.05% Igelal, 1.25% BSA, plus 1:200 FITC-anti-digoxigenin (Amersham) for dual-color experiments, for 35 minutes at 37°C. Excess detection mix was removed by washing the slides in 4 × SSC, 0.05% Igelal for 3 × 3 minutes. Slides were counterstained using Vectashield with DAPI (Vector Labs, Burlingame, CA, USA). Slides were viewed using an Olympus BX-61 epifluorescence microscope equipped with a cooled CCD camera and appropriate filters. Images were captured using SmartCapture 3 (Digital Scientific, Cambridge, UK).

We also performed cross-species FISH experiments to investigate whether MHC-containing chromosomes in the zebra finch correspond to chicken chromosome 16 (Additional file 7). These experiments involved co-hybridization of zebra finch BACs *TGAC-102M22* or *TGAC-86I22* in combination with chicken BAC *WAG65G9* (containing genetic markers *LEI0258* and *MCW0371*) to chicken and zebra finch chromosomes. Unfortunately, none of these experiments gave unequivocal evidence for colocalisation of chicken and zebra finch MHC BACs and therefore the data are not shown.

#### Gene and polymorphism survey via RFLP/Southern Blot

Restriction Fragment Length Polymorphisms (RFLP) were used to approximate the number of alleles for MHC Class I and Class IIB genes. We used the restriction enzyme Pvu II and digested seven micrograms of genomic DNA from ten captive zebra finches. These samples were run in two



identical parallel agarose gels that were transferred to nylon membranes and then hybridized with radioactively labeled zebra finch class I and II probes, respectively (for details on southern blot see Westerdahl et al. [44]). The probes were prepared as follows; An MHC class I/II PCR product was cloned into a bacterial vector (TOPO-TA cloning kit, Invitrogen, Carlsbad, CA, USA) inserts from five positive colonies were amplified and sequenced on a capillary sequencer according to manufacturer's protocol (Big Dye Terminator mix V3.1, Applied Biosystems, USA) and finally one MHC class I and one II insert, respectively, was amplified, cleaned and used as probes. The class I probe is a 271 bp exon 3 zebra finch DNA fragment (including primers), from a single colony, and it was amplified using the passerine Class I primers *PcaH1grw* (5' - TCC CCA CAG GTC TCC ACA CMA T - 3') and *A23H3* (5' - TTG CGC TCY AGC TCY YTC YGC C - 3') using standard PCR conditions. The zebra finch class II probe covers 207 bps in exon 2 and it was amplified, from a single colony, using the primers *2zffw1* (5' - TGT CAC TTC AYK AAC GGC ACG GAG - 3') and *2zfrv1* (5' - GTA GTT GTG CCG GCA GTA CGT GTC 3'). The probes were labelled with (α-32P)dCTP (PerkinElmer Boston, MA, USA) using the nick-translation technique (GE-healthcare, Little Chalfont, UK)

#### Comparative analysis of MHC genes

We estimated and compared gene density, across three avian lineages (chicken, quail and zebra finch) and the human MHC. To describe the human MHC, gene coordinates for protein coding genes were extracted from Ensembl [70] using the extended version of the human MHC map [79] as a template. To make an appropriate comparison among species, gene sets from human MHC were defined based on the flanking genes *SCGN* and *SYNGAP1*, but excluding pseudogenes, histones, tRNAs, vomeronasal and olfactory receptors. These genes were excluded because of their absence in the chicken MHC [20] and/or their lack of synteny with other MHC gene clusters in the zebra finch. For the chicken MHC gene set was based on the chicken extended MHC haplotype [20] and include the MHC-Y region. Gene coordinates were extracted using the latest annotation and assembly published in NCBI. For the quail MHC, extended regions have not been characterised and could not be included [35]. We calculated gene density by dividing the total number of genes by the total extent of the MHC region as defined above. Because zebra finch genes were often unmapped, and because of numerous pseudogenes, we approximated gene density for zebra finch using the two Class I BACs which appear to be a classical MHC region. Inclusion of all seven of the BACs also does not alter the conclusions of this analysis. Although the zebra finch MHC assembly remains fragmented, the patterns

revealed by this analysis highlight marked differences between zebra finch and chicken.

#### Phylogenetic analysis and tests for selection

*Class I* (exon 3) and *Class II* (exons 2 and 3) sequences were downloaded from GenBank (Accession #s given in Figure 1). Nucleotide sequences were aligned using MUSCLE [80] and then imported into Se-Al [81] for manual verification. Sequences were translated into amino acids and then adjusted by eye. All phylogenetic analyses were done using MrBayes v 3.1.2 [82]. For *Class II* we analyzed the two exons separately (not shown) and in a combined analysis where the data were partitioned by exon, and models were fitted to each codon position independently. To determine an appropriate root for passerine MHC sequences we first conducted analyses across all birds (including raptors, galliforms, and shorebirds, not shown here). For use in tests of selection, we conducted further analyses using only passerine sequences. MrBayes was run for 2.4 million generations, with 400,000 generations discarded as burn-in. One thousand sampled trees were then used to generate consensus trees and posterior probabilities. Trees from MrBayes and sequence alignments were analyzed in PAML [49,50] to test for evidence of selection acting on sites in the alignments. We used CODEML and tested two pairs of models using likelihood ratio tests. We tested the M1 model of nearly neutral evolution versus the M2 model of positive selection. We also tested the M7 model with the M8 model in which  $\omega$  ( $d_N/d_S$ ) can be greater than one. Both of these tests are routinely used to test for the influence of positive selection. Bayes Empirical Bayes analyses was used to identify specific sites with  $\omega > 1$  [83]. We also constructed phylogenies using only chicken and zebra finch sequences, but including putative zebra pseudogenes that spanned the exons of interest.

#### Additional file 1: Overgo probes used for BAC library screening.

Overgo probes targeting five genes of the MHC. Two pairs of probes were designed for each gene using sequences from the zebra genome trace archive.

#### Additional file 2: Self-self BLAST analysis of six BAC assemblies

(Class II: A to D, Class I: E to F). These results highlight the repetitive nature of these genomic regions, and the challenges faced in assembly.

#### Additional file 3: Genes found by BAC sequencing.

Genes found by BAC sequencing and manual and automated gene prediction.

#### Additional file 4: Two-color FISH mapping of TAP2 and MHC Class I

BACs. Depicted is the only case in which BACs putatively containing TAP2 and Class I colocalised. Colocalisation was on the W chromosome.

#### Additional file 5: BACs used in two color FISH mapping.

BACs used in dual-color FISH experiments with zebra finch MHC. These BACs are specific for zebra finch microchromosomes 9-15 and 17-28. None of these BACs cohybridized with MHC BACs. Because the whole genome assembly places some MHC genes on chromosome 22, we tested two chromosome 22 BACs. Both of these cohybridize with each other, and neither cohybridized with MHC BACs.

**Additional file 6: Phylogenetic analysis and selection on MHC Class I sequences.** **A)** Phylogenetic relationships among passerine MHC Class I, exon 3 sequences. Only one sequence with open reading frames were found in the zebra finch genome. The remaining sequences are from GenBank. **B)** Predicted amino acid sequences of the genomic sequence and one EST for MHC Class I. Stars represent sites showing evidence of selection in passerine birds. Note the similarity in the selected sites between raptors and passerines, both of which correspond well with the human PBR.

**Additional file 7: Preparation of chicken chromosomes.** The method for the preparation of chicken chromosome spreads is described.

## Abbreviations

BAC: bacterial artificial chromosome; Cds: coding sequence; ERV: endogenous retrovirus; EST: expressed sequence TAG; FISH: fluorescence in situ hybridization; LTR: long terminal repeat; RFLP: restriction fragment length polymorphism; UTR: untranslated region.

## Acknowledgements

Zebra finch chromosome preparations were kindly provided by Yuichiro Itoh. Zebra finch samples for RFLP analysis were provided by Maria Sandell. Tobias Lenz, Nancy Rotzel, Dan Janes and Miguel Alcaide provided assistance with lab work. Jim Kaufman, Dan Janes, Chris Organ and three anonymous reviewers gave helpful feedback on the manuscript. CNB was supported in part by NIH training grant T32DC006612 and by funds from Harvard University. RE was funded by a Marie Curie fellowship from the European Commission and HW was funded by the Swedish Research Council.

## Author details

<sup>1</sup>Department of Organismic & Evolutionary Biology, Museum of Comparative Zoology, Harvard University, Cambridge, MA 02138, USA. <sup>2</sup>Department of Animal & Plant Sciences, University of Sheffield, Sheffield, UK. <sup>3</sup>Department of Population Biology and Conservation Biology, Uppsala University, Uppsala, Sweden. <sup>4</sup>Department of Biosciences, University of Kent, Kent, UK. <sup>5</sup>Department of Animal Ecology, Lund University, Lund, Sweden. <sup>6</sup>School of Medicine, Genome Sequencing Center, Washington University, St Louis, MO, USA. <sup>7</sup>Roslin Institute, Division of Genetics & Genomics, University of Edinburgh, Edinburgh, UK. <sup>8</sup>Current address: Institute for Genomic Biology, University of Illinois at Urbana-Champaign, 1206 West Gregory Drive, Urbana, IL, USA.

## Authors' contributions

CNB and SVE designed the study. CNB screened the BAC library, annotated BAC clones and conducted phylogenetic analyses. RE and RG assessed the genome assembly. HK, TG and WW sequenced and assembled the BAC clones. RE and RG assisted with the annotation of BAC clones. HW conducted the RFLP analysis. MV and DG conducted the FISH mapping. DB sequenced plasmids used in the BAC assemblies. All authors contributed to the writing and/or the editing of the manuscript.

Received: 11 November 2009 Accepted: 1 April 2010

Published: 1 April 2010

## References

1. Zinkernagel RM, Doherty PC: MHC-restricted cytotoxic T cells: studies on the biological role of polymorphic major transplantation antigens determining T-cell restriction-specificity, function and responsiveness. *Adv Immunol* 1979, **27**:252-277.
2. Hughes AL, Nei M: Nucleotide substitution at major histocompatibility complex class-II loci - evidence for overdominant selection. *Proc Nat Acad Sci USA* 1989, **86**:958-962.
3. Hughes AL, Nei M: Pattern of nucleotide substitution at major histocompatibility complex class-I loci reveals overdominant selection. *Nature* 1988, **335**:167-170.
4. Bernatchez L, Landry C: MHC studies in nonmodel vertebrates: what have we learned about natural selection in 15 years? *J Evol Biol* 2003, **16**:363-377.
5. Roberts SC: Complexity and context of MHC-correlated mating preferences in wild populations. *Mol Ecol* 2009, **18**:3121-3123.
6. Ekblom R, Saether SA, Jacobsson P, Fiske P, Sahlman T, Grahn M, Kalas JA, Hoglund J: Spatial pattern of MHC class II variation in the great snipe (*Gallinago media*). *Mol Ecol* 2007, **16**:1439-1451.
7. Westerdaal H: Passerine MHC: genetic variation and disease resistance in the wild. *J Ornith* 2007, **148**:S469-S477.
8. Hale ML, Verduijn MH, Moller AP, Wolff K, Petrie M: Is the peacock's train an honest signal of genetic quality at the major histocompatibility complex? *J Evol Biol* 2009, **22**:1284-1294.
9. Von Schantz T, Wittzell H, Goransson G, Grahn M: Mate choice, male condition-dependent ornamentation and MHC in the pheasant. *Hereditas* 1997, **127**:133-140.
10. Kulski JK, Shiina T, Anzai T, Kohara S, Inoko H: Comparative genomic analysis of the MHC: the evolution of class I duplication blocks, diversity and complexity from shark to man. *Immunol Rev* 2002, **190**:95-122.
11. Ohta Y, Goetz W, Hossain MZ, Nonaka M, Flajnik MF: Ancestral organization of the MHC revealed in the amphibian *Xenopus*. *J Immunol* 2006, **176**:3674-3685.
12. Lukacs MF, Harstad H, Grimholt U, Beetz-Sargent M, Cooper GA, Reid L, Bakke HG, Phillips RB, Miller KM, Davidson WS, Koop BF: Genomic organization of duplicated major histocompatibility complex class I regions in Atlantic salmon (*Salmo salar*). *BMC Genomics* 2007, **8**:16.
13. Brigl M, Brenner MB: CD1: Antigen presentation and T cell function. *Ann Rev Immunol* 2004, **22**:817-890.
14. Martin LH, Calabi F, Milstein C: Isolation of CD1 genes - a family of major histocompatibility complex-related differentiation antigens. *Proc Nat Acad Sci USA* 1986, **83**:9154-9158.
15. Miller MM, Wang C, Parisini E, Coletta RD, Goto RM, Lee SY, Barral DC, Townes M, Roura-Mir C, Ford HL, Brenner MB, Dascher CC: Characterization of two avian MHC-like genes reveals an ancient origin of the CD1 family. *Proc Nat Acad Sci USA* 2005, **102**:8674-8679.
16. Salomonsen J, Sorensen MR, Marston DA, Rogers SL, Collen T, van Hateren A, Smith AL, Beal RK, Skjott K, Kaufman J: Two CD1 genes map to the chicken MHC, indicating that CD1 genes are ancient and likely to have been present in the primordial MHC. *Proc Nat Acad Sci USA* 2005, **102**:8668-8673.
17. Trowsdale J, Hanson I, Mockridge I, Beck S, Townsend A, Kelley A: Sequences encoded in the class II region of the MHC related to the 'ABC' superfamily of transporters. *Nature* 1990, **348**:741-744.
18. Bahram S, Arnold D, Bresnahan M, Strominger JL, Spies T: Two putative subunits of a peptide pump encoded in the human major histocompatibility complex class II region. *Proc Nat Acad Sci USA* 1991, **88**:10094-10098.
19. Kaufman J, Milne S, Gobel TWF, Walker BA, Jacob JP, Auffray C, Zoorob R, Beck S: The chicken B locus is a minimal essential major histocompatibility complex. *Nature* 1999, **401**:923-925.
20. Shiina T, Briles WE, Goto RM, Hosomichi K, Yanagiya K, Shimizu S, Inoko H, Miller MM: Extended gene map reveals tripartite motif, C-type lectin, and Ig superfamily type genes within a subregion of the chicken MHC-B affecting infectious disease. *J Immunol* 2007, **178**:7162-7172.
21. Briles WE, Goto RM, Auffray C, Miller MM: A polymorphic system related to but genetically independent of the chicken major histocompatibility complex. *Immunogenetics* 1993, **37**:408-414.
22. Miller MM, Goto R, Bernot A, Zoorob R, Auffray C, Bumstead N, Briles WE: 2 MHC Class-II genes map to the chicken RFP-Y system outside the B-complex. *Proc Nat Acad Sci USA* 1994, **91**:4397-4401.
23. Kaufman J, Volk H, Wallny HJ: A Minimal-Essential-Mhc and an Unrecognized-Mhc - 2 Extremes in Selection for Polymorphism. *Immunol Rev* 1995, **143**:63-88.
24. Kaufman J: Co-evolving genes in MHC haplotypes: the "rule" for nonmammalian vertebrates? *Immunogenetics* 1999, **50**:228-236.
25. Delany ME, Robinson CM, Goto RM, Miller MM: Architecture and organization of chicken microchromosome 16: order of the NOR, MHC-Y, and MHC-B subregions. *J Hered* 2009, **100**:507-514.
26. Kaufman J, Jacob J, Shaw I, Walker B, Milne S, Beck S, Salomonsen J: Gene organisation determines evolution of function in the chicken MHC. *Immunol Rev* 1999, **167**:101-117.
27. Jacob JP, Milne S, Beck S, Kaufman J: The major and a minor class II beta-chain (B-LB) gene flank the Tapasin gene in the B-F/B-L region of the

- chicken major histocompatibility complex. *Immunogenetics* 2000, **51**:138-147.
28. Dawkins R, Leelayuwat C, Gauderi S, Tay G, Hui J, Cattley S, Martinez P, Kulski J: **Genomics of the major histocompatibility complex: haplotypes, duplication, retroviruses and disease.** *Immunol Rev* 1999, **167**:275-304.
29. Cole RK: **Studies on the genetic resistance to Marek's disease.** *Avian Diseases* 1968, **12**:9-28.
30. Plachy J, Pink JRL, Hala K: **Biology of the chicken MHC (B complex).** *Crit Rev Immunol* 1992, **12**:47-49.
31. Gregory TR: **A bird's-eye view of the C-value enigma: genome size, cell size, and metabolic rate in the class aves.** *Evolution* 2002, **56**:121-130.
32. Organ CL, Shedlock AM, Meade A, Pagel M, Edwards SV: **Origin of avian genome size and structure in non-avian dinosaurs.** *Nature* 2007, **446**:180-184.
33. Brown JW, Rest JS, Garcia-Moreno J, Sorenson MD, Mindell DP: **Strong mitochondrial DNA support for a Cretaceous origin of modern avian lineages.** *BMC Biology* 2008, **6**:6.
34. Shiina T, Shimizu S, Hosomichi K, Kohara S, Watanabe S, Hanzawa K, Beck S, Kulski JK, Inoko H: **Comparative genomic analysis of two avian (quail and chicken) MHC regions.** *J Immunol* 2004, **172**:6751-6763.
35. Hosomichi K, Shiina T, Suzuki S, Tanaka M, Shimizu S, Iwamoto S, Hara H, Yoshida Y, Kulski JK, Inoko H, Hanzawa K: **The major histocompatibility complex (Mhc) class IIB region has greater genomic structural flexibility and diversity in the quail than the chicken.** *BMC Genomics* 2006, **7**.
36. Chaves LD, Krueh SB, Reed KM: **Defining the Turkey MHC: Sequence and Genes of the B Locus.** *J Immunol* 2009, **183**:6530-6537.
37. Chaves LD, Krueh SB, Reed KM: **Characterization of the Turkey MHC chromosome through genetic and physical mapping.** *Cytogenet Gen Res* 2007, **117**:213-220.
38. Edwards SV, Wakeland EK, Potts WK: **Contrasting histories of avian and mammalian Mhc genes revealed by class II B sequences from songbirds.** *Proc Nat Acad Sci USA* 1995, **92**:12200-12204.
39. Wittzell H, Madsen T, Westerdahl H, Shine R, von Schantz T: **MHC variation in birds and reptiles.** *Genetica* 1998, **104**:301-309.
40. Westerdahl H, Wittzell H, von Schantz T: **Mhc diversity in two passerine birds: no evidence for a minimal essential Mhc.** *Immunogenetics* 2000, **52**:92-100.
41. Hess CM, Gasper J, Hoekstra HE, Hill CE, Edwards SV: **MHC class II pseudogene and genomic signature of a 32-kb cosmid in the house finch (*Carpodacus mexicanus*).** *Genome Research* 2000, **10**:613-623.
42. Edwards SV, Gasper J, March M: **Genomics and polymorphism of Agph-DAB1, an Mhc class II B gene in red-winged blackbirds (*Agelaius phoeniceus*).** *Mol Biol Evol* 1998, **15**:236-250.
43. Edwards SV, Gasper J, Garrigan D, Martindale D, Koop BF: **A 39-kb sequence around a blackbird Mhc class II gene: Ghost of selection past and songbird genome architecture.** *Mol Biol Evol* 2000, **17**:1384-1395.
44. Westerdahl H, Wittzell H, von Schantz T: **Polymorphism and transcription of Mhc class I genes in a passerine bird, the great reed warbler.** *Immunogenetics* 1999, **49**:158-170.
45. Warren WC, Clayton DF, Ellegren H, Arnold AP, Hillier LW, Kunster A, Searle S, White S, Vilella AJ, Fairly S, Heger A, Kong L, Ponting CP, Jarvis ED, Mello CV, Minx P, Lovell P, Velho TAF, Ferris M, Balakrishnan CN, Sinha S, Blatti C, London SE, Li Y, Lin Y, George J, Sweedler J, Southey B, Gunaratne P, Watson M, Nam K, Backstrom N, Smeds L, Nabholz B, Itoh Y, Whitney O, Pfenning AR, Howard J, Volker M, Skinner BM, Griffin DK, Ye L, McLaren WM, Flicek P, Quesada V, Velasco G, Lopez-Otin C, Puente XS, Olender T, Lancet D, Smit AFA, Hubley R, Konkel MK, Walker JA, Batzer MA, Gu W, Pollock DD, Chen L, Cheng Z, Eichler EE, Stapley J, Slate J, Ekblom R, Birkhead T, Burke T, Burt D, Scharff C, Adam I, Hughes R, Sultan M, Soldatov A, Lehrach H, Edwards SV, Yang S, Li X, Graves R, Fulton L, Nelson J, Chinwalla A, Hou S, Mardis ER, Wilson RK: **The genome of a songbird.** *Nature*.
46. Ekblom R, Balakrishnan CN, Burke T, Slate J: **Digital gene expression analysis of the zebra finch genome.** *BMC Genomics*.
47. Ovcharenko I, Loots GG, Hardison RC, Miller W, Stubbs L: **zPicture: Dynamic alignment and visualization tool for analyzing conservation profiles.** *Genome Res* 2004, **14**:472-477.
48. Gasper JS, Shiina T, Inoko H, Edwards SV: **Songbird genomics: Analysis of 45 kb upstream of a polymorphic Mhc class II gene in red-winged blackbirds (*Agelaius phoeniceus*).** *Genomics* 2001, **75**:26-34.
49. Yang ZH: **PAML 4: Phylogenetic analysis by maximum likelihood.** *Mol Biol Evol* 2007, **24**:1586-1591.
50. Yang ZH: **PAML: a program package for phylogenetic analysis by maximum likelihood.** *Comp App Biosci* 1997, **13**:555-556.
51. Brown JH, Jardetzky TS, Gorga JC, Stern LJ, Urban RG, Strominger JL, Wiley DC: **3-Dimensional structure of the human class-II histocompatibility antigen HLA-DR1.** *Nature* 1993, **364**:33-39.
52. Bjorkman PJ, Saper MA, Samraoui B, Bennett WS, Strominger JL, Wiley DC: **The foreign antigen binding site and T cell recognition regions of class I histocompatibility antigens.** *Nature* 1987, **329**:593-602.
53. Alcaide M, Edwards SV, Cadahia L, Negro JJ: **MHC class I genes of bird of prey: isolation, polymorphism, and diversifying selection.** *Cons Genet*.
54. Alcaide M, Edwards SV, Negro JJ: **Characterization, polymorphism, and evolution of MHC class IIB genes in birds of prey.** *J Mol Evol* 2007, **65**:541-554.
55. Deakin JE, Siddle HV, Cross JG, Belov K, Graves JA: **Class I genes have split from the MHC in the tammar wallaby.** *Cytogenet Genome Res* 2007, **116**:205-211.
56. Siddle HV, Deakin JE, Coghill P, Hart E, Cheng YY, Wong ESW, Harrow J, Beck S, Belov K: **MHC-linked and un-linked class I genes in the wallaby.** *BMC Genomics* 2009, **10**:310.
57. Kaufman J: **The simple chicken major histocompatibility complex: life and death in the face of pathogens and vaccines.** *Phil Trans Roy Soc Lond B* 2000, **355**:1077-1084.
58. Aguilar A, Smith TB, Wayne RK: **A comparison of variation between a MHC pseudogene and microsatellite loci of the little greenbul (*Andropadus virens*).** *BMC Evol Biol* 2005, **5**:47.
59. Miller HC, Lambert DM: **Gene duplication and gene conversion in class II MHC genes of New Zealand robins (*Petroicidae*).** *Immunogenetics* 2004, **56**:178-191.
60. Burri R, Niculita-Hirzel H, Roulin A, Fumagalli L: **Isolation and characterization of major histocompatibility complex (MHC) class IIB genes in the barn owl (*Aves: Tyto alba*).** *Immunogenetics* 2008, **60**:543-550.
61. Hosomichi K, Miller MM, Goto RM, Wang YJ, Suzuki S, Kulski JK, Nishibori M, Inoko H, Hanzawa K, Shiina T: **Contribution of mutation, recombination, and gene conversion to chicken Mhc-B haplotype diversity.** *J Immunol* 2008, **181**:3393-3399.
62. Hughes CR, Miles S, Walbroehl JM: **Support for the minimal essential MHC hypothesis: a parrot with a single, highly polymorphic MHC class IIB gene.** *Immunogenetics* 2008, **60**:219-231.
63. Ekblom R, Grahm M, Hoglund J: **Patterns of polymorphism in the MHC class II of a non-passerine bird, the great snipe (*Gallinago media*).** *Immunogenetics* 2003, **54**:734-741.
64. Hackett SJ, Kimball RT, Reddy S, Bowie RC, Braun EL, Braun MJ, Chojnowski JL, Cox WA, Han KL, Harshman J, Huddleston CJ, Marks BD, Miglia KJ, Moore WS, Sheldon FH, Steadman DW, Witt CC, Yuri T: **A phylogenomic study of birds reveals their evolutionary history.** *Science* 2008, **320**:1763-1768.
65. Moon DA, Veniamin SM, Parks-Dely JA, Magor KE: **The MHC of the duck (*Anas platyrhynchos*) contains five differentially expressed class I genes.** *J Immunol* 2005, **175**:6702-6712.
66. Eddy S: **Profile hidden Markov models.** *Bioinformatics* 1998, **14**:755-763.
67. Birney E, Clamp M, Durbin R: **GeneWise and genomewise.** *Genome Res* 2004, **14**:988-995.
68. Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL-W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.** *Nuc Acid Res* 1994, **22**:4673-4680.
69. Hall T: **BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT.** *Nuc Acid Symp Ser* 1999, **95**:98.
70. Birney E, Clamp M, Kraspyk A, Slater G, Hubbard T, Curwen V, Stabenau A, Stupka E, Huminiecki L, Potter S: **Ensembl: A multi-genome computational platform.** *Am J Hum Genet* 2001, **69**:219.
71. Edwards SV, Grahm M, Potts WK: **Dynamics of Mhc evolution in birds and crocodilians: amplification of class II genes with degenerate primers.** *Mol Ecol* 1995, **4**:719-729.
72. Shedlock AM, Janes DE, Edwards SV: **Amniote phylogenomics: testing evolutionary hypotheses with BAC library scanning and targeted clone analysis of large-scale DNA sequences from reptiles.** *Phylogenomics* Totowa, NJ: Humana Press; Murphy W 2008.

73. Ross MT, LaBrie S, McPherson J, Stanton VP: *Current Protocols in Human Genetics* New York: WileyDracopoli NC 1999, 561-565.
74. de la Blastide M, McCombie WR: **Assembling genomic DNA sequences with PHRAP.** *Current Protocols in Bioinformatics* 2007, Supp. 17, Unit 11.4..
75. Salomov AA, Solovyev VV: **Ab initio gene finding in Drosophila genomic DNA.** *Gen Res* 2000, **11**:817-832.
76. Burge C, Karlin S: **Prediction of complete gene structures in human genomic DNA.** *J Mol Biol* 1997, **268**:78-94.
77. Smit AFA, Hubley R, Green P: **RepeatMasker Open 3.0.** 1996 [<http://www.repeatmasker.org>].
78. Itoh Y, Arnold AP: **Chromosomal polymorphism and comparative painting analysis in the zebra finch.** *Chromosome Res* 2005, **13**:47-56.
79. Horton R, Wilming L, Rand V, Lovering RC, Bruford EA, Khodiyar VK, Lush MJ, Povey S, Talbot CC Jr, Wright MW, Wain HM, Trowsdale J, Ziegler A, Beck S: **Gene map of the extended human MHC.** *Nat Rev Genet* 2004, **5**:889-899.
80. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput.** *Nuc Acid Res* 2004, **32**:1792-1797.
81. Rambaut A: **Se-Al: Sequence Alignment Editor.** 1996 [<http://evolve.zoo.ac.uk>].
82. Huelsenbeck JP, Ronquist F: **MrBayes: Bayesian inference of phylogenetic trees.** *Bioinformatics* 2001, **17**:754-755.
83. Yang ZH, Wong WSW, Nielsen R: **Bayes empirical Bayes inference of amino acid sites under positive selection.** *Mol Biol Evol* 2005, **22**:1107-1118.

doi:10.1186/1741-7007-8-29

**Cite this article as:** Balakrishnan *et al.*: Gene duplication and fragmentation in the zebra finch major histocompatibility complex. *BMC Biology* 2010 **8**:29.

**Submit your next manuscript to BioMed Central and take full advantage of:**

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at  
[www.biomedcentral.com/submit](http://www.biomedcentral.com/submit)

